

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Rediscovery of Good-Turing estimators via Bayesian nonparametrics

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/1591184> since 2016-09-03T11:55:58Z

Published version:

DOI:10.1111/biom.12366

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

SUMMARY. The problem of estimating discovery probabilities originated in the context of statistical ecology. In recent years it has become popular due to its frequent appearance in challenging applications arising in genetics, linguistics, designs of experiments, machine learning, etc. A full range of statistical approaches, parametric and nonparametric, as well as frequentist and Bayesian, has been proposed for estimating discovery probabilities. In this article we investigate the relationships between the celebrated Good–Turing approach, which is a frequentist nonparametric approach introduced in the 1940s, and a Bayesian nonparametric approach recently introduced in the literature. Specifically, under the assumption of a two parameter Poisson–Dirichlet prior, we show that Bayesian nonparametric estimators of discovery probabilities are asymptotically equivalent, for a large sample size, to suitably smoothed Good–Turing estimators. As a by-product, we introduce and investigate a methodology for deriving exact and asymptotic credible intervals for discovery probabilities with the Bayesian nonparametric estimators of discovery probabilities. The proposed methodology is illustrated by a comprehensive simulation study and the analysis of Expressed Sequence Tags data generated by sequencing a complementary DNA library.

KEY WORDS: Asymptotic equivalence; Bayesian nonparametrics; Credible intervals; Discovery probabilities; Expressed Sequence Tags; Good–Toulmin estimator; Good–Turing estimator; Smoothing technique; Two parameter Poisson–Dirichlet prior.

1. Introduction

Consider a population of individuals $(X_i)_{i \geq 1}$ belonging to an (ideally) infinite number of species $(X_i^*)_{i \geq 1}$ with unknown proportions $(p_i)_{i \geq 1}$. Given an initial observed sample of size n , a quantity of practical interest is the probability $D_{n,m}(l)$ of observing at the $(n+m+1)$ -th drawn a species with frequency $l \geq 0$ in the enlarged sample of size $n+m$, with the additional sample being unobserved. Formally, if $N_{i,n+m}$ denotes the frequency of X_i^* in the enlarged sample, then

$$D_{n,m}(l) = \sum_{i \geq 1} p_i \mathbb{1}_{\{l\}}(N_{i,n+m}). \quad (1)$$

Clearly $D_{n,m}(0)$ corresponds to the proportion of yet unobserved species or, equivalently, the probability of discovering a new species at the $(n+m+1)$ -th drawn. The random probability (1) is typically referred to as the $(m; l)$ -discovery. While the $(0; l)$ -discovery is of interest for estimating the probability of discovering new species or rare species, the $(m; l)$ -discovery is typically of interest in decision problems regarding the size of the additional sample to collect.

A full range of statistical approaches, parametric and nonparametric as well as frequentist and Bayesian, have been proposed for estimating $D_{n,m}(l)$. These approaches have originally found applications in ecology, and their importance has grown considerably in recent years, driven by challenging applications arising in genetics, bioinformatics, linguistics, designs of experiments, machine learning, etc. See Bunge and Fitzpatrick (1993) and Bunge et al. (2014) for comprehensive re-

views. In this article we investigate the relationships between two approaches for estimating $D_{n,m}(l)$: (i) the Good–Turing approach which appeared in the literature in Good (1953), and first developed by Alan Turing and J. Good during their collaboration at Bletchley Park in the 1940s; (ii) the Bayesian nonparametric approach recently introduced by Lijoi et al. (2007) and Favaro et al. (2014). In order to state our main contributions, we first review the relevant aspects of these two nonparametric approaches.

1.1. The Good–Turing Approach

Let \mathcal{H} be a parametric statistical hypothesis for the species composition that is \mathcal{H} determines the species composition by specifying a distribution function γ on \mathbb{N} and with a finite number of unknown parameters. Let (X_1, \dots, X_n) be a random sample from \mathcal{H} , and let by $M_{l,n}$ the number of species with frequency l in the sample. According to Good (1953), an estimator of $D_{n,0}(l)$ is $(l+1)\mathbb{E}_{\mathcal{H}}[M_{l+1,n+1}]/(n+1)$, where $\mathbb{E}_{\mathcal{H}}$ denotes the expected value with respect to the distribution function γ of \mathcal{H} . For any $m \geq 1$ let us consider the enlarged sample $(X_{n+1}, \dots, X_{n+m})$, and define γ_m according to Good and Toulmin (1956), an estimator of $D_{n,m}(0)$ is $\sum_{i \geq 1} (-\gamma)^{i-1} i \mathbb{E}_{\mathcal{H}}[M_{i,n+m}]/n$. Note that, in general, $\mathbb{E}_{\mathcal{H}}[M_{l+1,n+1}]$ and $\mathbb{E}_{\mathcal{H}}[M_{i,n+m}]$ do not depend on the observed sample, unless the parameters characterizing γ are estimated using such a sample. Several exact and asymptotic results have been thoroughly discussed in Good (1953) and, more recently, in Favaro et al. (2014). We mention the Zipf-type distributions and the Dirichlet process distributions.

The large n approximate estimator (2) is known as the Good–Turing estimator. A similar large n approximation was proposed in Good and Toulmin (1956) for $\check{D}_{n,m}(0; \mathcal{H})$. Specifically,

$$\check{D}_{n,m}(0; \mathcal{H}) \simeq \check{D}_{n,m}(0) = \frac{1}{n} \sum_{i \geq 1} (-\gamma)^{i-1} i m_{i,n}. \quad (3)$$

$\check{D}_{n,m}(0)$ is known as the Good–Toulmin estimator for the $(m; 0)$ -discovery. As observed by Good and Toulmin (1956), due to the alternating sign of the series which appears in the estimator (3), if γ is large then $\check{D}_{n,m}(0)$ can yield inadmissible estimates. This instability arises even for values of m moderately larger than n , typically m greater than n is enough for it to appear.

A peculiar feature of $\check{D}_{n,0}(l)$ is that it depends on $m_{l+1,n}$, and not on $m_{l,n}$ as one would intuitively expect for an estimator of the $(0; l)$ -discovery. Such a feature, combined with the irregular behavior of the $m_{l,n}$ ’s for large l , makes $\check{D}_{n,0}(l)$ a sensible approximation only if l is sufficiently small with respect to n . Indeed for some large l one might observe that $m_{l,n} > 0$ and $m_{l+1,n} = 0$, which provides the absurd estimate $\check{D}_{n,0}(l) = 0$, or that $m_{l,n} < m_{l+1,n}$ although the overall observed trend for $m_{l,n}$ is to decrease as l increases. In order to overcome these drawbacks Good (1953) suggested to smooth the irregular series of $m_{l,n}$ ’s into a more regular series to be used as a proxy. If $m'_{l,n}$ ’s are the smoothed $m_{l,n}$ ’s with respect to a smoothing rule \mathcal{S} , then $\check{D}_{n,0}(l; \mathcal{S}) = (l+1)m'_{l+1,n}/n$ is a more accurate approximation than $\check{D}_{n,0}(l)$. Common smoothing rules consider $m'_{l,n}$, as a function of l , to be approximately parabolic or, alternatively, $m'_{l,n}$ to be a certain proportion of the number of species in \mathbf{X}_n . An alternative method assumes \mathcal{H} to be selected from a superpopulation with an assigned distribution. This flexible method was hinted at in Good (1953) and then left as an open problem.

1.2. The Bayesian Nonparametric Approach

The approach in Lijoi et al. (2007) and Favaro et al. (2012) is based on the randomization of p_i ’s. This is somehow reminiscent of the superpopulation smoothing hinted at by Good (1953). Specifically, let $P = \sum_{i \geq 1} p_i \delta_{X_i^*}$ be a discrete random probability measure, namely $(p_i)_{i \geq 1}$ are nonnegative random weights such that $\sum_{i \geq 1} p_i = 1$ almost surely, and $(X_i^*)_{i \geq 1}$ are random locations independent of $(p_i)_{i \geq 1}$ and independent and identically distributed as a nonatomic distribution. The sample \mathbf{X}_n is drawn from a population with species composition determined by P , i.e.,

$$X_i | P \stackrel{\text{iid}}{\sim} P \quad i = 1, \dots, n \quad P \sim \mathcal{P}, \quad (4)$$

a random probability measure with respect to a squared loss prior, Lijoi et al. (2007) and Favaro et al. (2012) derived a Bayesian nonparametric estimator for the $(0; l)$ -discovery. Specifically, let \mathbf{X}_n be a sample of size n from P , and let $K_n = k_n$ species with corresponding frequency counts $(M_{1,n}, \dots, M_{n,n}) = (m_{1,n}, \dots, m_{n,n})$. From Lijoi et al. (2007), the Bayesian nonparametric estimator $\hat{D}_{n,m}(0)$, with respect to a squared loss

$$\hat{D}_{n,m}(0) = \frac{\theta + \sigma k_n}{\theta + n} \frac{(\theta + n)}{(\theta + n)}$$

for any $m \geq 0$, where $(a)_n = \prod_{0 \leq i \leq n-1} (a - i)$ and $(a)_0 = 1$. For any $m \geq 0$, let $(X_{n+1}, \dots, X_{n+m})$ be a conditional unobserved sample from $P_{\alpha, \theta}$. According to Favaro et al. (2012), the Bayesian nonparametric estimator of $D_{n,m}(l)$, with respect to a squared loss

$$\begin{aligned} \hat{D}_{n,m}(l) &= \sum_{i=1}^l \binom{m}{l-i} m_{i,n} (i - \sigma)_{l+1-i} \\ &\quad + (1 - \sigma)_l \binom{m}{l} (\theta + \sigma k_n) \frac{(\theta)}{(\theta + n)} \end{aligned}$$

for any $l = 1, \dots, n + m$. According to the results in (5) and (6), the Bayesian nonparametric estimator has several advantages with respect to the Good–Turing estimator: (i) it leads directly to exact estimators, without the need of large n approximations; (ii) $\hat{D}_{n,0}(l)$ depends on $m_{l,n}$, and not on $m_{l+1,n}$, thus avoiding the irregular behavior of the $m_{l,n}$ ’s.

1.3. Contributions of the Paper and Conclusions

Let $a_n \simeq b_n$ mean that $\lim_{n \rightarrow \infty} a_n/b_n = 1$. In this article we show that the Bayesian nonparametric estimator $\hat{D}_{n,0}(l)$ is asymptotically equivalent to the Good–Turing estimator $\check{D}_{n,0}(l)$ as n tends to infinity, to a Good–Turing estimator based on smoothed frequency counts. More precisely, we show that $\hat{D}_{n,0}(l) \simeq \check{D}_{n,0}(l; \mathcal{S}_{PD})$ as $n \rightarrow \infty$ for any smoothing rule such that $m_{l,n}$ is smooth

$$m'_{l,n} = \frac{\sigma(1 - \sigma)_{l-1}}{l!} k_n$$

While smoothing techniques were introduced in Good (1953) as an ad hoc tool for post processing of the Good–Turing estimator to improve the performance of $\check{D}_{n,0}(l)$, our

Good–Turing estimator $\mathcal{D}_{m,0}(l)$ in which $m_{l+1,m}$ is replaced by a smoothed version, via \mathcal{S}_{PD} , of the Bayesian nonparametric estimator $\hat{\mathcal{M}}_{n,m}(l+1)$ of the number of species with frequency l in the enlarged sample. As a by-product of this result, we introduce a methodology for deriving large m asymptotic credible intervals for $\hat{\mathcal{D}}_{n,m}(l)$, thus completing the study in Lijoi et al. (2007) and Favaro et al. (2012). While the $\text{PD}(\sigma, \theta)$ prior leads to an explicit expression for the posterior distribution of $\mathcal{D}_{n,m}(l)$, this expression involves combinatorial coefficients whose evaluation for large m is cumbersome, thus preventing its implementation for determining exact credible intervals. Our methodology thus provides a fundamental tool in many situations of practical interest, arising especially in genomics, where m is required to be very large and only a small portion of the population is sampled.

Our results are illustrated through a simulation study and the analysis of Expressed Sequence Tags (ESTs) data generated by sequencing a benchmark complementary DNA (cDNA) library. By means of a simulation study, we compare $\hat{\mathcal{D}}_{n,0}(l; \mathcal{S}_{\text{PD}})$ with smoothed Good–Turing estimators obtained by using the Poisson smoothing and a smoothing technique in Sampson (2001). Simulated data are generated from the Zeta distribution, whose power law behavior is common in numerous applications. In order to detect the effects of different smoothing techniques, we compare the smoothed Good–Turing estimators with $\hat{\mathcal{D}}_{n,0}(l)$ and $\hat{\mathcal{D}}_{n,0}(l)$. A second numerical illustration is devoted to the large m asymptotic credible intervals for the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,m}(l)$. Using ESTs data, we compare asymptotic confidence intervals for the Good–Toulmin estimator $\hat{\mathcal{D}}_{n,m}(0)$ with asymptotic credible intervals for its Bayesian nonparametric counterpart $\hat{\mathcal{D}}_{n,m}(0)$. This study completes the numerical illustration presented in Favaro et al. (2009) and Favaro et al. (2012) on the same ESTs data.

In Section 2, we present and discuss the asymptotic equivalence between the Good–Turing approach and the Bayesian nonparametric approach under the assumption of the $\text{PD}(\sigma, \theta)$ prior. As a by-product of this asymptotic analysis, in Section 3 we introduce a methodology for associating large m asymptotic credible intervals to $\mathcal{D}_{n,m}(l)$. Section 4 contains numerical illustrations. Proofs of our results, as well as related additional materials, and the Matlab code for computing asymptotic credible intervals are available as web-based supplementary materials.

2. Good Turing Estimators via Bayesian Nonparametrics

Under a $\text{PD}(\sigma, \theta)$ prior, the most notable difference between the Good–Turing estimator and its Bayesian nonparametric counterpart can be traced back to the different use of the in-

$\mathcal{D}_{n,m}(l)$ and $\mathcal{D}_{m,0}(l)$. With a slight abuse of notation, throughout this section we write $X|Y$ to denote a random variable whose distribution coincides with the conditional distribution of X given Y .

2.1. Large n Asymptotic Equivalences for $\hat{\mathcal{D}}_{n,0}(l)$

We start by recalling the predictive distribution obtained by maximizing $P_{\sigma,\theta}$. Let \mathbf{X}_n be a sample of size n containing k_n species $X_1^*, \dots, X_{k_n}^*$ with frequencies $(n_{1,n}, \dots, n_{k_n,n})$. According to the de Finetti theorem, \mathbf{X}_n is part of an exchangeable sequence. The predictive distribution has been characterized by Pitman (1972) and

$$\mathbb{P}[\mathbf{X}_{n+1} \in \cdot | \mathbf{X}_n] = \frac{\theta + \sigma k_n}{\theta + n} v_0(\cdot) + \frac{1}{\theta + n} \sum_{i=1}^{k_n} \frac{n_{i,n}}{n} \delta_{X_i^*}(\cdot)$$

with v_0 being a nonatomic probability measure. The predictive probability (8) is referred to as the predictive distribution of $P_{\sigma,\theta}$. Note that $\hat{\mathcal{D}}_{n,0}(l)$ can be recovered from (5) and (6) one has $\hat{\mathcal{D}}_{n,0}(0) = (\theta + \sigma)/(\theta + n)$ and $\hat{\mathcal{D}}_{n,0}(l) = (l - \sigma)m_{l,n}/(\theta + n)$, respectively. See for details on (8), and on the joint distribution $(N_{1,n}, \dots, N_{k_n,n})$ induced by (8).

The asymptotic equivalence between $\hat{\mathcal{D}}_{n,0}(l; \mathcal{S}_{\text{PD}})$ and $\hat{\mathcal{D}}_{n,0}(l)$ relies on an interesting interplay between the large n asymptotic behaviors of K_n and $\text{PD}(\sigma, \theta)$ prior. Specifically, let $A_n \stackrel{\text{a.s.}}{\simeq} B_n$ as $n \rightarrow +\infty$ if $\lim_{n \rightarrow +\infty} A_n/B_n = 1$ almost surely, namely A_n and B_n are almost surely asymptotically equivalent as $n \rightarrow +\infty$. By a direct application of Theorem 3.8 and Pitman (2006), one obtains the asymptotic

$$M_{l,n} \stackrel{\text{a.s.}}{\simeq} \frac{\sigma(1 - \sigma)_{l-1}}{l!} K_n$$

as $n \rightarrow +\infty$. In other terms, under a $\text{PD}(\sigma, \theta)$ prior, as the sample size n tends to infinity the number of species with frequency l becomes a proportion $\sigma(1 - \sigma)_{l-1}$ of the total number of species. We refer to the web appendix for details on (9). The next theorem combines (8) and (9) to establish the asymptotic equivalence between $\hat{\mathcal{D}}_{n,0}(l; \mathcal{S}_{\text{PD}})$ and $\hat{\mathcal{D}}_{n,0}(l)$.

THEOREM 1. *Let \mathbf{X}_n be a sample of size n featuring $K_n = k_n$ species with corresponding frequencies $(n_{1,n}, \dots, n_{k_n,n})$. Then, as $n \rightarrow +\infty$,*

the frequency count $m_{l,n}$ by taking the proportion $\sigma(1 - \sigma)_{l-1}/l!$ of k_n . Such a smoothing rule is somehow related to the Poisson smoothing \mathcal{S}_{Poi} , originally introduced by Good (1953), in which the frequency count $m_{l,n}$ is approximately equal to a proportion $e^{-\lambda}\lambda^{\tau+l-1}/(\tau+l-1)!$ of k_n , for any $\lambda > 0$ and $\tau \geq 0$ such that $\sum_{l \geq 0} \hat{D}_{n,0}(l; \mathcal{S}_{\text{Poi}}) = 1$. See Chapter 2 in Engen (1978) for a common example of Poisson smoothing where $\tau = 1$ and $\lambda = n/k_n$. In particular \mathcal{S}_{PD} is related to the Poisson smoothing corresponding to the choice $\tau = 0$ and to a suitable randomization of the parameter λ . Specifically, let us denote by P_λ a discrete random variable with distribution $\mathbb{P}[P_\lambda = l] = e^{-\lambda}\lambda^{l-1}/(l-1)!$, that is the Poisson smoothing with $\tau = 0$ and $\lambda > 0$. If $G_{a,b}$ is Gamma random variable with parameter (a, b) and L_σ is a discrete random variable with distribution $\mathbb{P}[L_\sigma = l] = \sigma(1 - \sigma)_{l-1}/l!$, then according to Devroye (1993) $L_\sigma \stackrel{d}{=} 1 + P_{G_{1,1}G_{1,1-\sigma}/G_{1,\sigma}}$ where $G_{1,1}$, $G_{1,1-\sigma}$ and $G_{1,\sigma}$ are mutually independent.

A peculiar feature of the smoothing rule \mathcal{S}_{PD} is that it depends only on $\sigma \in (0, 1)$. This is because \mathcal{S}_{PD} is obtained by suitably combining (9), which does not depend of the parameter θ , with other two large n asymptotic equivalences independent of θ , namely: (i) $\hat{D}_{n,0}(0) \simeq \sigma k_n/n$ and ii) $\hat{D}_{n,0}(l) \simeq (l - \sigma)m_{l,n}/n$. We conjecture that these asymptotic equivalences, as well as (9), hold for a more general class of priors considered in Lijoi et al. (2007) and Favaro et al. (2012). This is the class of Gibbs-type priors introduced by Pitman (2003) and including two of the most commonly used nonparametric priors, i.e., the $\text{PD}(\sigma, \theta)$ prior and the normalized generalized Gamma prior. See De Blasi et al. (2015) for details. In other terms, our conjecture is that Theorem 1 holds for any Gibbs-type prior, that is the smoothing rule \mathcal{S}_{PD} is invariant with respect to the choice of any prior in the Gibbs class. Intuitively, different smoothing rules for different Gibbs-type priors, if they exist, necessarily require to investigate the high-order large n asymptotic behavior of $\hat{D}_{n,0}(l)$, and then combine it with a corresponding refinement of the asymptotic equivalence in (9). Work on this is ongoing.

2.2. Large m Asymptotic Equivalences for $\hat{D}_{n,m}(l)$

Let \mathbf{X}_n be a sample of size n from $P_{\alpha,\theta}$ featuring $K_n = k_n$ species with frequency counts $(M_{1,n}, \dots, M_{n,n}) = (m_{1,n}, \dots, m_{n,n})$. For any $m \geq 1$ let $(X_{n+1}, \dots, X_{n+m})$ be an additional unobserved sample. Let $K_m^{(n)}$ be the number of new species in $(X_{n+1}, \dots, X_{n+m})$ and let $M_{l,m}^{(n)}$ denote the number of species with frequency l in (X_1, \dots, X_{n+m}) . Since the additional sample is assumed to be not observed, let us introduce a randomized version of $\hat{D}_{n+m,0}(0)$ and $\hat{D}_{n+m,0}(l)$ as

$$D_{0,m}^{(n)} = \frac{\theta + \sigma k_n + \sigma K_m^{(n)}}{\theta + n + m} \quad (11)$$

according to the expression (6), $(K_n, M_{1,n}, \dots, M_{n,n})$ is a sufficient statistic for $\hat{D}_{n,m}(l)$ and, therefore $D_{n,m}^{(n)}(l) | \mathbf{X}_n$ takes on the interpretation of the conditional distribution, with respect to \mathbf{X}_n , of the $(m; l)$ -discovery. By means of the identities introduced in (11) and (12) the distribution of $D_{0,m}^{(n)} | \mathbf{X}_n$ and $D_{n,m}^{(n)}(l) | \mathbf{X}_n$ can be obtained from the distribution of $K_m^{(n)} | \mathbf{X}_n$ and $M_{l,m}^{(n)} | \mathbf{X}_n$, respectively. These distributions have been obtained in Lijoi et al. (2007) and Favaro et al. (2012). See the web appendix for details on the derivation. In particular, Proposition 1 in Favaro et al. (2012) states that

$$\hat{K}_{n,m} = \mathbb{E}[K_m^{(n)} | \mathbf{X}_n] = \frac{(\theta/\sigma + k_n)}{(\theta + n)_m} ((\theta + \sigma k_n) + \sigma k_n)$$

which is the Bayesian nonparametric estimator of the expectation, to a squared loss function, of $K_m^{(n)}$. Furthermore, Proposition 7 in Favaro et al. (2012) states that

$$\begin{aligned} \hat{\mathcal{M}}_{n,m}(l) &= \mathbb{E}[M_{l,m}^{(n)} | \mathbf{X}_n] \\ &= \sum_{i=1}^l \binom{m}{l-i} m_{i,n} (i - \sigma)_{l-i-1} \\ &\quad + (1 - \sigma)_{l-1} \binom{m}{l} (\theta + \sigma k_n) \end{aligned}$$

which is the Bayesian nonparametric estimator of the expectation, to a squared loss function, of $M_{l,m}^{(n)}$. By means of (11) and (12) one obtains $\hat{D}_{n,m}(0) \simeq (\theta + \sigma k_n + \sigma \hat{K}_{n,m})/(\theta + n + m)$ and $\hat{D}_{n,m}(l) \simeq (l - \sigma)\hat{\mathcal{M}}_{n,m}(l)/(\theta + n + m)$, which provide a representation for the estimators of the $(m; 0)$ - and $(m; l)$ -discovery, respectively.

Similarly to Theorem 1, an asymptotic equivalence for $\hat{D}_{n,m}(l)$ and $\hat{D}_{m,0}(l)$ relies on the interpretation of the m asymptotic behaviors of the random variables $K_m^{(n)} | \mathbf{X}_n$ and $M_{l,m}^{(n)} | \mathbf{X}_n$. Specifically, for any $n \geq 1$, by means of Proposition 2 in Favaro et al. (2009) and Proposition 1 in Gnedin et al. (2007) one obtains the following asymptotic equivalence

$$M_{l,m}^{(n)} | \mathbf{X}_n \stackrel{\text{a.s.}}{\simeq} \frac{\sigma(1 - \sigma)_{l-1}}{l!} K_m^{(n)}$$

as $m \rightarrow +\infty$. In other terms, under a $P_{\alpha,\theta}$ prior, the m asymptotic equivalence between $M_{l,m}^{(n)}$ and $K_m^{(n)}$ coincides with the large n asymptotic equivalence between $\hat{D}_{n,m}(l)$ and $\hat{D}_{m,0}(l)$. We refer to the web appendix for details.

$$\hat{D}_{n,m}(l) \simeq (l+1) \frac{\mathcal{M}_{n,m}(l+1)}{m} \simeq (l+1) \frac{\frac{(l+1)!}{l!} \mathcal{K}_{n,m}}{m}. \quad (14)$$

Besides discovery probabilities one is also interested in cumulative discovery probabilities, which are generalizations of the $(m; l)$ -discovery defined as follows. For any $\tau \geq 1$, let $\{l_1, \dots, l_\tau\}$ be a collection of distinct indexes such that $l_i \in \{0, 1, \dots, n+m\}$ for any $i = 1, \dots, \tau$. We define the $(m; l_1, \dots, l_\tau)$ -discovery as the cumulative discovery probability $D_{n,m}(l_1, \dots, l_\tau) = \sum_{1 \leq i \leq \tau} D_{n,m}(l_i)$. Hence, the Bayesian nonparametric estimator of $(m; l_1, \dots, l_\tau)$ -discovery is

$$\hat{D}_{n,m}(l_1, \dots, l_\tau) = \sum_{i=1}^{\tau} \hat{D}_{n,m}(l_i).$$

Such a generalization of the $(m; l)$ -discovery is mainly motivated by several applications of practical interest in which one aims at estimating the probability of discovering the so-called rare species. Specifically, these are species not yet observed or observed with a frequency smaller than a certain threshold τ . Of course, large n and large m asymptotic equivalences for the estimator $\hat{D}_{n,m}(l_1, \dots, l_\tau)$ follow by a direct application of Theorem 1 and Theorem 2, respectively.

3. Credible Intervals for $\hat{D}_{n,m}(l_1, \dots, l_\tau)$

While deriving the estimator $\hat{D}_{n,m}(l)$, Lijoi et al. (2007) and Favaro et al. (2012) did not consider the problem of associating a measure of uncertainty to $\hat{D}_{n,m}(l)$. Such a problem reduces to the problem of evaluating the distribution of $D_{l,m}^{(n)} | \mathbf{X}_n$ by combining (11) and (12) with the distributions of $K_m^{(n)} | \mathbf{X}_n$ and $M_{l,m}^{(n)} | \mathbf{X}_n$ recalled in the web appendix. While the distribution of $D_{l,m}^{(n)} | \mathbf{X}_n$ is explicit, in many situations of practical interest the additional sample size m is required to be very large and the computational burden for evaluating this posterior distribution becomes overwhelming. This happens, for instance, in various genomic applications where one has to deal with relevant portions of cDNA libraries which typically consist of millions of genes. In this section, we show how to exploit the large m asymptotic behavior of $D_{l,m}^{(n)} | \mathbf{X}_n$ in order to associate asymptotic credible intervals to the estimator $\hat{D}_{n,m}(l)$.

Let \mathbf{X}_n be a sample from $P_{\alpha,\theta}$ featuring $K_n = k_n$ species $X_1^*, \dots, X_{K_n}^*$ with frequencies $(N_{1,n}, \dots, N_{K_n,n}) = (n_{1,n}, \dots, n_{k_n,n})$. Let $Z_{\alpha,\theta,k_n}^{(n)} \stackrel{d}{=} B_{k_n+\theta/\alpha, n/\sigma-k_n} Z_{\alpha,(\theta+n)/\sigma}$ where $B_{a,b}$ is a Beta random variable with parameter (a, b) and $Z_{\alpha,q}$ has density function $f_{Z_{\alpha,q}}(z) = \Gamma(q\sigma + 1) z^{q-1-1/\sigma} f_\sigma(z^{-1/\sigma}) / \sigma \Gamma(q +$

$D_{(l_1, \dots, l_\tau), m}^{(n)} = \sum_{1 \leq i \leq \tau} D_{l_i, m}^{(n)}$. The distribution takes on the interpretation of the posterior $(m; l_1, \dots, l_\tau)$ -discovery. In the next proposition we apply the fluctuation limit (15) to the cumulative probability $D_{(l_1, \dots, l_\tau), m}^{(n)} | \mathbf{X}_n$.

PROPOSITION 1. *Let \mathbf{X}_n be a sample of size n from $P_{\alpha,\theta}$ featuring $K_n = k_n$ species with corresponding frequencies $(M_{1,n}, \dots, M_{k_n,n}) = (m_{1,n}, \dots, m_{k_n,n})$. Then, as $m \rightarrow +\infty$, we have*

$$\frac{D_{(l_1, \dots, l_\tau), m}^{(n)}}{m^{\sigma-1}} | \mathbf{X}_n \xrightarrow{w} \left(\sum_{i=1}^{\tau} \frac{\sigma(1-\sigma)_{l_i}}{l_i!} \right).$$

Fluctuation limits (15) and (16) provide us with a good approximation of the distribution of $D_{l,m}^{(n)} | \mathbf{X}_n$ and $D_{(l_1, \dots, l_\tau), m}^{(n)} | \mathbf{X}_n$. The same fluctuation limits hold for any $\sigma > 0$ such that, as $m \rightarrow +\infty$, $r(m) \simeq m^{\sigma-1}$. This produces a scaling factor finer than $m^{\sigma-1}$. Indeed, it is easily verified that, as soon as θ and n are not smaller than m ,

$$\hat{D}_{n,m}'(l) = m^{\sigma-1} \frac{\sigma(1-\sigma)_l}{l!} \mathbb{E}[Z_{\alpha,\theta,k_n}^{(n)}]$$

with $\mathbb{E}[Z_{\alpha,\theta,k_n}^{(n)}] = (k_n + \theta/\sigma) \Gamma(\theta + n) / \Gamma(\theta + n + \sigma)$ from $\hat{D}_{n,m}(l)$. Hence, the corresponding asymptotic credible intervals could be far from the exact ones. Of course, the same issue appears for the estimator $\hat{D}_{n,m}(l_1, \dots, l_\tau)$. For this reason we consider the factors $r^*(m, l)$ and $r^*(m, l_1, \dots, l_\tau)$ in such a way that $\hat{D}_{n,m}(l) = r^*(m, l) (\sigma(1-\sigma)_l / l!) \mathbb{E}[Z_{\alpha,\theta,k_n}^{(n)}]$ and $\hat{D}_{n,m}(l_1, \dots, l_\tau) = r^*(m, l_1, \dots, l_\tau) \sum_{1 \leq i \leq \tau} (\sigma(1-\sigma)_{l_i} / l_i!) \mathbb{E}[Z_{\alpha,\theta,k_n}^{(n)}]$.

$$\hat{D}_{n,m}^*(l) = r^*(m, l) \frac{\sigma(1-\sigma)_l}{l!} \mathbb{E}[Z_{\alpha,\theta,k_n}^{(n)}]$$

and

$$\hat{D}_{n,m}^*(l_1, \dots, l_\tau) = r^*(m, l_1, \dots, l_\tau) \left(\sum_{i=1}^{\tau} \frac{\sigma(1-\sigma)_{l_i}}{l_i!} \right) \mathbb{E}[Z_{\alpha,\theta,k_n}^{(n)}]$$

It can be easily verified that, as $m \rightarrow +\infty$, $r^*(m, l_1, \dots, l_\tau) \simeq m^{\sigma-1}$. Explicit expressions for the factors $r^*(m, l)$ and $r^*(m, l_1, \dots, l_\tau)$ are given in the web appendix. The reader is referred to Favaro et al. (2012) for more details.

$(r^*(m, l)\sigma(1 - \sigma)_{lS_1}/l!, r^*(m, l)\sigma(1 - \sigma)_{lS_2}/l!)$ is a 95% asymptotic credible interval for $\hat{D}_{n,m}(l)$. Analogous observations hold true for the estimator $\hat{D}_{n,m}(l_1, \dots, l_r)$. In order to determine the quantiles s_1 and s_2 , we resort to a simulation algorithm for sampling the limiting random variable $Z_{\alpha, \theta, k_n}^{(n)}$. Note that, according to the definition of $Z_{\alpha, \theta, k_n}^{(n)}$, this procedure involves sampling from the random variable $Z_{\alpha, q}$ with density function $f_{Z_{\alpha, q}}(z) = \Gamma(q\sigma + 1)z^{q-1-1/\sigma} f_\sigma(z^{-1/\sigma})/\sigma\Gamma(q + 1)$.

A strategy for sampling $Z_{\alpha, q}$ was proposed by Favaro et al. (2009). Specifically, let $L_{\alpha, q} = Z_{\alpha, q}^{-1/\sigma}$ and we introduce a Gamma random variable U_q with parameter $(q, 1)$. Then, conditionally on $U_q = u$, the distribution of $L_{\alpha, q}$ has density function proportional to $f_\sigma(x) \exp\{-ux\}$. Therefore, the problem of sampling from $Z_{\alpha, q}$ boils down to the problem of sampling from an exponentially tilted stable distribution. Here we improve the sampling scheme proposed in Favaro et al. (2009) by resorting to the fast rejection algorithm recently proposed in Hofert (2011) for sampling from an exponentially tilted positive σ -stable random variable. Summarizing, in order to generate random variates from the distribution of $Z_{\alpha, \theta, k_n}^{(n)}$, we have the following steps: (i) sample $B_{k_n + \theta/\sigma, n/\sigma - k_n}$; (ii) sample $G_{(\theta+n)/\sigma, 1}$ and set $U_{(\theta+n)/\sigma} = G_{(\theta+n)/\sigma, 1}^{1/\sigma}$; (iii) given $U_{(\theta+n)/\sigma} = u$, sample $L_{\alpha, (\theta+n)/\sigma}$ from density proportional to $f_\sigma(x) \exp\{-ux\}$, by means of the fast rejection sampling, and set $Z_{\alpha, (\theta+n)/\sigma} = L_{\alpha, (\theta+n)/\sigma}^{-\sigma}$; (iv) set $Z_{\alpha, \theta, k_n}^{(n)} = B_{k_n + \theta/\sigma, n/\sigma - k_n} Z_{\alpha, (\theta+n)/\sigma}$.

4. Illustrations

In order to implement our results, the first issue to be faced is the specification of the parameter (σ, θ) in the $\text{PD}(\sigma, \theta)$ prior. Hereafter, following the approach of Lijoi et al. (2007) and Favaro et al. (2012), we resort to an empirical Bayes procedure. Specifically let \mathbf{X}_n be a sample from $P_{\sigma, \theta}$ featuring $K_n = k_n$ species with frequencies $(N_{1,n}, \dots, N_{K_n,n}) = (n_{1,n}, \dots, n_{k_n,n})$. The empirical Bayes procedure consists in choosing θ and σ that maximize the distribution of \mathbf{X}_n . This, under a $\text{PD}(\sigma, \theta)$ prior, corresponds to setting $(\sigma, \theta) = (\hat{\sigma}, \hat{\theta})$, where

$$(\hat{\sigma}, \hat{\theta}) = \arg \max_{(\sigma, \theta)} \left\{ \frac{\prod_{i=0}^{k_n-1} (\theta + i\sigma)}{(\theta)_n} \prod_{i=1}^{k_n} (1 - \sigma)_{n_{i,n}-1} \right\}. \quad (18)$$

One could also specify a prior distribution on the parameter (σ, θ) and then seek a full Bayesian inference. However, in terms of estimating $D_{n,m}(l)$, there are no relevant differences between this fully Bayes approach and the empirical Bayes approach, given the posterior distribution of (σ, θ) is highly concentrated; this is typically the case of large datasets since

We compare the performance of the (0; l)-discovery with responding Good–Turing estimators and Bayesian nonparametric estimators, for some choices of l . We draw 500 samples of size $n = 1000$ from $\text{PD}(\sigma, \theta)$ with scale parameter $s = 1.5$. Recall that the random variable Z is such that $\mathbb{P}[Z = z] = z^{-s}/C(s)$ for $s > 1$. Next we order the samples according to the number of observed distinct species k_n and we select the first 5 samples. Specifically, for $i = 1, 2, \dots, 5$, the i -th sample is composed by 100 samples featuring k_n observed distinct species k_n that stays below the order $(i - 1)/5$ and $i/5$ of the empirical distribution. We therefore pick at random one sample from each of the 5 samples with the corresponding index i . This procedure yields a total number of 5 samples of 1000 observed distinct species compositions.

We use these simulated datasets to compare the performance of the estimators for the (0; l)-discovery with $D_{n,0}(l)$, for $l = 0, 1, 5, 10, 20, 30$. Specifically, we compare the Bayesian nonparametric estimator $\hat{D}_{n,0}(l)$, the smoothed Good–Turing estimator $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{PD}})$, and the Poisson smoothed Good–Turing estimator $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{Poi}})$ with $\tau = 1$ and $\lambda = n/k_n$. We consider the so-called Simple Good–Turing estimator $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$, which is a popular smoothed estimator discussed in Chapter 7 of Sørensen (1992). Specifically, in the Simple Good–Turing estimator, the rule \mathcal{S}_{SGT} consists in first computing, for each species i , the values $z_{i,n}$ that take into account both the observed counts $m_{i,n}$ and the surrounding zero values. Then, we resort to a line of best fit for the pairs $(m_{i,n}, z_{i,n})$ in order to obtain the smoothed values $\hat{z}_{i,n}$.

Table 1 summarizes the result of our simulations. As an overall measure for the performance of the estimators, we use the sum of squared error $\text{SSE}(\hat{D}(l))$ of the generic estimator $\hat{D}(l)$ of the (0, l)-discovery, defined as $\sum_{0 \leq l \leq n} (\hat{D}(l) - d_{n,0}(l))^2$, with $d_{n,0}(l)$ being the true value of $D_{n,0}(l)$. By looking at the SSE in Table 1, we see that $\hat{D}_{n,0}(l)$ and $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$ are much more accurate than the others. As expected, the Good–Turing estimator shows good performance only for small values of l . Large values of l arise for large frequencies thus explaining the high values of the resulting SSE. For instance, since the number of species that has frequency $l = 20$ is small, the frequency $l = 21$, the Good–Turing estimator $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$ is 0 while, clearly, there is positive probability that a species appeared 20 times in the sample. The estimator $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{Poi}})$ yields a smaller SSE than $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{PD}})$. The accuracy of $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{PD}})$ and $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{Poi}})$ and $\tilde{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$, shows that the pa-

	$\hat{D}_{n,0}(l)$	0.0871	0.0939	0.1004	0.1016
$l = 0$	$\check{D}_{n,0}(l)$	0.0870	0.0950	0.1040	0.1040
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{Poi}})$	0.0006	0.0008	0.0008	0.0011
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{PD}})$	0.0859	0.0933	0.1002	0.1011
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$	0.0870	0.0950	0.1040	0.1040
	$D_{n,0}(l)$	0.0273	0.0272	0.0478	0.0365
	$\hat{D}_{n,0}(l)$	0.0320	0.0312	0.0301	0.0319
$l = 1$	$\check{D}_{n,0}(l)$	0.0320	0.0220	0.0160	0.0240
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{Poi}})$	0.0047	0.0054	0.0059	0.0073
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{PD}})$	0.0316	0.0307	0.0290	0.0311
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$	0.0319	0.0221	0.0161	0.0240
	$D_{n,0}(l)$	0.0060	0.0238	0.0132	0.0154
	$\hat{D}_{n,0}(l)$	0.0044	0.0173	0.0086	0.0215
$l = 5$	$\check{D}_{n,0}(l)$	0.0240	0.0180	0.0120	0.0180
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{Poi}})$	0.1148	0.1206	0.1243	0.1332
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{PD}})$	0.0126	0.0114	0.0101	0.0111
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$	0.0044	0.0176	0.0089	0.0219
	$D_{n,0}(l)$	0.0105	0	0.0105	0.0092
	$\hat{D}_{n,0}(l)$	0.0094	0	0.0093	0.0093
$l = 10$	$\check{D}_{n,0}(l)$	0	0	0.0220	0.0110
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{Poi}})$	0.0816	0.0769	0.0738	0.0664
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{PD}})$	0.0082	0.0072	0.0062	0.0070
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$	0.0093	0	0.0094	0.0093
	$D_{n,0}(l)$	0	0.0142	0.0169	0
	$\hat{D}_{n,0}(l)$	0	0.0193	0.0193	0
$l = 20$	$\check{D}_{n,0}(l)$	0	0	0	0
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{Poi}})$	0.0001	0.0000	0.0000	0.0000
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{PD}})$	0.0053	0.0046	0.0038	0.0043
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$	0	0.0194	0.0195	0
	$D_{n,0}(l)$	0.0260	0	0	0
	$\hat{D}_{n,0}(l)$	0.0293	0	0	0
$l = 30$	$\check{D}_{n,0}(l)$	0	0	0	0
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{Poi}})$	0.0000	0.0000	0.0000	0.0000
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{PD}})$	0.0041	0.0035	0.0029	0.0033
	$\check{D}_{n,0}(l; \mathcal{S}_{\text{SGT}})$	0.0292	0	0	0
	$\text{MSE}(\hat{D}_{n,0})$	0.0006	0.0016	0.0007	0.0007
	$\text{MSE}(\check{D}_{n,0})$	0.3475	0.3773	0.3460	0.3575
	$\text{MSE}(\check{D}_{n,0}(\mathcal{S}_{\text{Poi}}))$	0.2657	0.2723	0.2765	0.2769
	$\text{MSE}(\check{D}_{n,0}(\mathcal{S}_{\text{PD}}))$	0.1748	0.1748	0.1753	0.1746
	$\text{MSE}(\check{D}_{n,0}(\mathcal{S}_{\text{SGT}}))$	0.0007	0.0018	0.0014	0.0008

underlying the smoothing rules \mathcal{S}_{Poi} and \mathcal{S}_{PD} are not suitable for data generated according to a Zeta distribution.

4.2. Credible Intervals for $\hat{D}_{n,m}(l_1, \dots, l_\tau)$

We illustrate the implementation of the asymptotic credible intervals for the Bayesian nonparametric estimator

$\hat{D}_{n,m}(l_1, \dots, l_\tau)$ through the analysis of EST by sequencing a benchmark cDNA library, an efficient way to characterize expressed organism. The rate of gene discovery depends redundancy of the cDNA library from which are obtained. Correctly estimating the relative

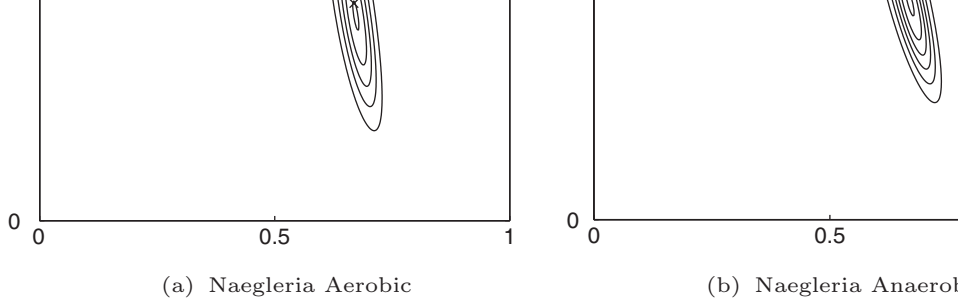


Figure 1. Contour lines of the posterior distribution of the parameter (σ, θ) . The cross marks denote the estimates obtained by means of the empirical Bayes procedure (18).

such libraries, as well as other quantities such as the probability of sampling a new or a rarely observed gene, is of fundamental importance since it allows one to optimize the use of expensive experimental sampling techniques. Hereafter, we consider the *Naegleria gruberi* cDNA libraries prepared from cells grown under different culture conditions, namely aerobic and anaerobic. See Susko and Roger (2004) for additional details.

The *Naegleria gruberi* aerobic library consists of $n = 959$ ESTs with $k_n = 473$ distinct genes and $m_{i,959} = 346, 57, 19, 12, 9, 5, 4, 2, 4, 5, 4, 1, 1, 1, 1, 1$, for $i = \{1, 2, \dots, 12\} \cup \{16, 17, 18\} \cup \{27\} \cup \{55\}$. The *Naegleria gruberi* anaerobic library consists of $n = 969$ ESTs with $k_n = 631$ distinct genes and $m_{i,969} = 491, 72, 30, 9, 13, 5, 3, 1, 2, 0, 1, 0, 1$, for $i \in \{1, 2, \dots, 13\}$. A fully Bayesian approach involves the specification of a prior distribution for the parameter (σ, θ) . Let us consider independent priors for σ and θ , namely a Uniform distribution on $(0, 1)$ for σ and a Gamma distribution with shape parameter 1 and scale parameter 100, for θ . Figure 1 shows the contour lines of the posterior distribution of (σ, θ) ; note that these posterior distributions are rather concentrated on a small range of values for σ . The empirical Bayes approach (18) lead to the following estimates for $(\hat{\sigma}, \hat{\theta})$: $(\hat{\sigma}, \hat{\theta}) = (0.669, 46.241)$ for the *Naegleria gruberi* aerobic library and $(\hat{\sigma}, \hat{\theta}) = (0.656, 155.408)$ for the *Naegleria gruberi* anaerobic library. These values are very close to the mode of the corresponding posterior distributions. See the cross marks in Figure 1. As a matter of fact, the fully Bayesian approach and the empirical Bayes approach lead to very similar estimates for $D_{n,m}(l)$. For instance, by adopting both the empirical Bayes approach and the fully Bayesian approach we get $\hat{D}_{n,0}(0) = 0.36$ for the *Naegleria gruberi* aerobic library and $\hat{D}_{n,0}(0) = 0.51$ for the *Naegleria gruberi* anaerobic library. This observation supports our choice of undertaking the empirical Bayes approach (18). The reader is referred to the web appendix for a sensitivity analysis of

the asymptotic credible intervals for $\hat{D}_{n,m}(0)$, the choice of the parameter (σ, θ) .

We now focus on the *Naegleria gruberi* aerobic library and observe that the estimates of the exact estimator $\hat{D}_{n,m}(0)$, are 0.289, 0.165, 0.080, respectively, while the estimates provided by the asymptotic estimator are 0.367, 0.171, 0.080. It is apparent that the estimates that are close to the exact estimator are very large. This motivates the use of $\hat{D}_{n,m}^*(0)$ with a more accurate scaling factor. The asymptotic intervals hold for the *Naegleria gruberi* anaerobic library. A comparative study between the asymptotic intervals and $\hat{D}_{n,m}^*(0)$, as well as the corresponding empirical Bayes estimates, is presented in Table 2.

The estimator $\hat{D}_{n,m}(0)$ is compared with the exact estimator $\hat{D}_{n,m}(0)$. Confidence intervals have been devised in Mao (2004) for the exact approach, and asymptotic credible intervals are also compared. We focus on $m \in [0, n]$, the fact that $\hat{D}_{n,m}(0)$ is known to be small for small m , namely $m \leq n$. See Mao (2004) highlights common features for the exact and the *Naegleria gruberi* libraries. When m increases, both approaches provide similar estimates for $\hat{D}_{n,m}(0)$. However, even for small values of m , the asymptotic intervals are narrower than the corresponding 95% confidence intervals. This difference is substantial when m increases. While the exact intervals show a regular behavior around the point estimates, with intervals that tend to increase slowly, estimates obtained with the Gaussian approach and corresponding confidence intervals show a similar behavior. The latter approach can lead to very different behaviors, as m approaches

$(n = 999)$	$10n$	0.169	0.171	(0.168, 0.164)	0.169
<i>Naegleria</i> Anaerobic ($n = 969$)	n	0.409	0.533	(0.074, 0.086)	0.080
	$10n$	0.232	0.241	(0.229, 0.254)	0.232
	$100n$	0.109	0.109	(0.103, 0.115)	0.109

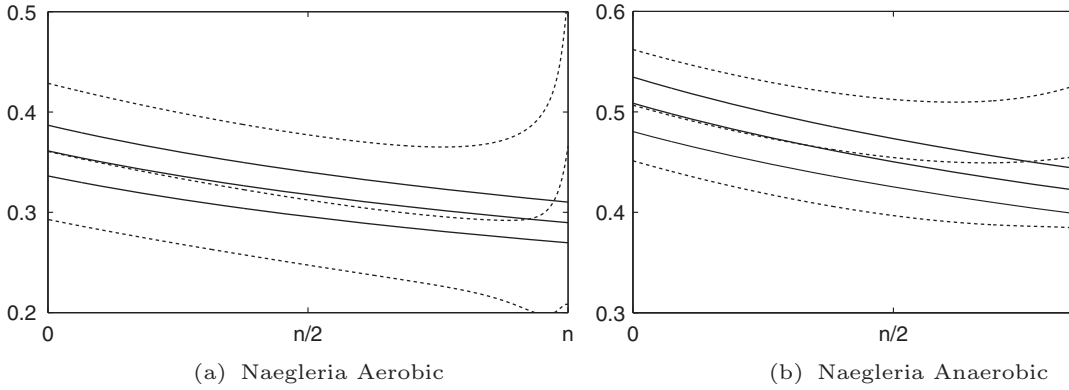


Figure 2. Comparison of Good-Toulmin estimator $\hat{D}_{n,m}(0)$ (inner dashed curves) and Bayesian nonparametric estimator $\hat{D}_{n,m}(0)$ (inner solid curves) for m ranging in $[0, n]$. The Good-Toulmin estimates are endowed with 95% credible intervals (outer dashed curves). Bayesian nonparametric estimators are endowed with asymptotic 95% credible intervals (outer solid curves).

Table 3

Naegleria aerobic and Naegleria anaerobic libraries. $\hat{D}_{n,m}(l)$, for $l = 0, 1, 2, 3, 4$, and $\hat{D}_{n,m}(0, \dots, \tau)$, for τ corresponding asymptotic 95% credible intervals (c.i.)

	Library	$m = n$		$m = 2n$		estimate
		estimate	95% c.i.	estimate	95% c.i.	
$(m; 0)$ -discovery	aerobic	0.289	(0.267, 0.312)	0.253	(0.234, 0.273)	0.231
	anaerobic	0.409	(0.387, 0.431)	0.358	(0.339, 0.378)	0.326
$(m; 1)$ -discovery	aerobic	0.093	(0.084, 0.101)	0.083	(0.076, 0.089)	0.075
	anaerobic	0.130	(0.123, 0.137)	0.117	(0.111, 0.124)	0.108
$(m; 2)$ -discovery	aerobic	0.061	(0.057, 0.066)	0.054	(0.050, 0.059)	0.050
	anaerobic	0.080	(0.076, 0.085)	0.075	(0.071, 0.079)	0.070
$(m; 3)$ -discovery	aerobic	0.046	(0.042, 0.049)	0.041	(0.038, 0.045)	0.038
	anaerobic	0.059	(0.056, 0.062)	0.055	(0.052, 0.058)	0.052
$(m; 4)$ -discovery	aerobic	0.036	(0.033, 0.039)	0.034	(0.031, 0.036)	0.031
	anaerobic	0.045	(0.042, 0.047)	0.044	(0.042, 0.046)	0.042
$(m; 0, 1, 2, 3)$ -discovery	aerobic	0.490	(0.452, 0.528)	0.432	(0.399, 0.465)	0.394
	anaerobic	0.679	(0.642, 0.716)	0.606	(0.573, 0.640)	0.556
$(m; 0, 1, 2, 3, 4)$ -discovery	aerobic	0.526	(0.485, 0.563)	0.465	(0.430, 0.501)	0.425
	anaerobic	0.724	(0.685, 0.763)	0.650	(0.615, 0.686)	0.599
$(m; 0, 1, 2, 3, 4, 5)$ -discovery	aerobic	0.556	(0.514, 0.599)	0.494	(0.456, 0.532)	0.452
	anaerobic	0.760	(0.718, 0.801)	0.686	(0.649, 0.723)	0.634

tainty are provided for these estimates. In Table 3 we summarize estimates of the $(m; l)$ -discovery for $l = 0, \dots, 4$ and of the $(m; l_1, \dots, l_\tau)$ -discovery for $\tau = 3, 4, 5$. These estimates are endowed with asymptotic 95% credible intervals obtained by combining asymptotic results displayed in (15) and (16) with the choice of the scaling factors $r^*(m, l)$ and $r^*(m, l_1, \dots, l_\tau)$, respectively. Table 3 thus complete the illustrations presented in Favaro et al. (2009) and Favaro et al. (2012).

5. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2, 3 and 4 are available with this paper at the *Biometrics* website on Wiley Online Library. The Matlab code for computing the asymptotic credible intervals for $\hat{D}_{n,m}(l_1, \dots, l_\tau)$ is also available at the Biometrics website on Wiley Online Library.

ACKNOWLEDGEMENTS

The authors are grateful to an Associate Editor and an anonymous referee for their constructive comments and suggestions. Stefano Favaro is supported by the European Research Council through StG N-BNP 306406. Yee Whye Teh is supported by the European Research Council through the European Unions Seventh Framework Programme (FP7/2007-2013) ERC grant agreement 617411.

REFERENCES

- Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association* **88**, 364–373.
- Bunge, J., Willis, A., and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application* **1**, 427–445.
- De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans-*

- Favaro, S., Lijoi, A., and Prünster, I. (2012). Estimating the discovery probability. *Biometrika* **99**, 1108–1118.
- Favaro, S., Lijoi, A., and Prünster, I. (2013). Estimating the discovery probability for Gibbs-type exchangeable random partitions. *Applied Probability* **23**, 1721–1754.
- Gnedin, S., Hansen, B., and Pitman, J. (2011). The discovery problem with infinitely many species. *Probability Surveys* **8**, 1–26.
- Good, I. J. (1953). The population frequency of species: the estimation of population parameters. *Biometrika* **40**, 264.
- Good, I. J. and Toulmin, G. H. (1956). The discovery of new species and the increase in population coverage. *Biometrika* **43**, 45–63.
- Hofert, M. (2011). Efficiently sampling new species. *Computational Statistics & Data Analysis* **55**, 1108–1118.
- Lijoi, A., Mena, R. H., and Prünster, I. (2012). Nonparametric estimation of the probability of discovery of new species. *Biometrika* **94**, 769–786.
- Mao, C. X. (2004). Prediction of the conditional probability of discovery covering a new class. *Journal of the Royal Statistical Society A* **167**, 1108–1118.
- Pitman, J. (1995). Exchangeable and partitionable random partitions. *Probability Theory and Related Fields* **107**, 145–158.
- Pitman, J. (2003). Poisson-Kingman partitions. *Probability Theory: A Festschrift for Terry Speed*. *Lecture Notes Monograph Series* **40**, 1–22. Cambridge University Press, Cambridge, MA.
- Pitman, J. (2006). *Combinatorial Stochastic Processes*. *Progress in Probability* **16**. Birkhäuser, Basel.
- Sampson, G. (2001). *Empirical Linguistics*. Cambridge University Press, Cambridge, New York.
- Susko, E. and Roger, A. J. (2004). Estimating the number of species: rates of gene discovery and expression. *Journal of Molecular Evolution* **59**, 1–11.
- Susko, E. and Roger, A. J. (2004). Estimating the number of species: frequencies in EST surveys. *Bioinformatics* **20**, 1108–1118.

Received October 2014. Revised

Accepted June 2015.